(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－167098
（P2001－167098A）
(43)公開日　平成13年６月22日(2001.6.22)

(51)Int.Cl.'	識別記号	ＦＩ		ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３２０Ｚ	５Ｂ０７５
		15/40	３１０Ｃ	
			３８０Ａ	

審査請求　未請求　請求項の数４　ＯＬ　（全 11 頁）

(21)出願番号　特願平11－347026

(22)出願日　平成11年12月７日(1999.12.7)

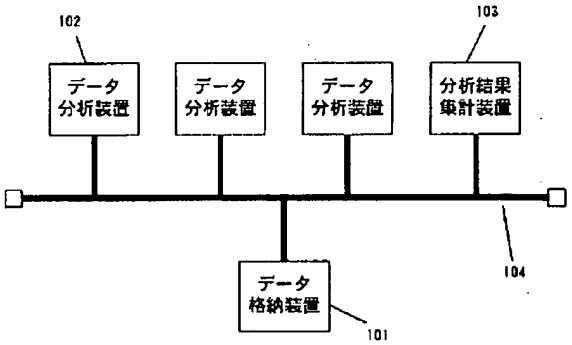
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　牧　秀行
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　森田　豊久
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　大量データの分散並列分析方法

(57)【要約】
【課題】大量のデータから知識を発見するデータマイニングの並列分散処理方法に関しては、従来、データを分割して複数の処理装置に分配するので、処理装置間で分析対象データの転送が必要となったり、並列分散処理方法を実行するためには、前処理としてデータベースから各処理装置へデータを分配する必要があるという問題がある。また、処理装置に大容量の主記憶が必要であるという問題がある。
【解決手段】単数、または複数のデータ格納手段、分析結果集計手段、複数のデータ分析手段を用いる。データ格納手段は分析対象データを一回送信し、複数のデータ分析手段がこれを受信する。各々のデータ分析手段は受信したデータを対象として分析を行い、その後、それぞれのデータ分析手段において得られた分析結果を分析結果集計手段が集計し、全体の分析結果とする。

図 1



1

【特許請求の範囲】

【請求項１】　単数、または複数のデータ格納手段に格納されたデータを対象とし、複数のデータ分析手段を用いるデータ分析において、前記の複数のデータ分析手段は前記のデータ格納手段から同一のデータを入力し、前記複数のデータ分析手段のそれぞれにおいて分析を行い、前記複数のデータ分析手段のそれぞれにおける分析結果をまとめて全体の分析結果をすることを特徴とするデータ分析方法。

【請求項２】　前記複数のデータ分析手段は前記のデータ格納手段が一回出力した分析対象データを共有することを特徴とする請求項１に記載のデータ分析方法。

【請求項３】　複数のデータ分析手段が共有する共有記憶手段を用い、前記複数のデータ分析手段は各々の分析結果を前記共有記憶手段へ出力することを特徴とする請求項１、および請求項２に記載のデータ分析方法。

【請求項４】　請求項１、２、および３に記載の並列分散分析方法を計算機で実行するための計算機プログラムを格納した記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、大量のデータを対象とするデータ分析技術に関する。

【０００２】

【従来の技術】大量のデータから知識を発見する技術はデータマイニングと呼ばれている。発見される知識の具体例として、相関ルール（Association Rule）がよく知られている。相関ルールの基本的概念は文献「Mining Association Rules between Setof Items in Large Databases」（proceeding of ACM SIGMOD、1993）に説明されている。それによれば、I_1 から I_m までの m 個の二値属性（アイテムと呼ばれ、0 か 1 の一方の値を持つ）と、アイテムに対応する m 個の要素からなる二値ベクトル（トランザクションと呼ばれる）と、このトランザクションの集合 T を考えた時、相関ルールは「X → I_j」と記述される。ここで、X は I_1から I_m までの m 個のアイテムのうちのいくつかからなるアイテムの集合（アイテムセットと呼ばれる）、I_j は X に含まれない単一のアイテムである。1つのトランザクション t と、1つのアイテム i を考えた時、t の要素のうち、i に対応するものの値が 1 であれば、トランザクション t は アイテム i を満足すると言い、t がアイテムセット X に含まれる全てのアイテムを満足する時、トランザクション t はアイテムセット X を満足すると言う。トランザクション集合 T において、アイテムセット X を満足するトランザクションの数を K、アイテムセット X を満足し、かつアイテム I_j をも満足するトランザクションの数を J とした時、割合 J/K を相関ルール「X → I_j」の「コンフィデンス」と呼ぶ。また、トランザクション集合 T の全体に対する上記 J の割合を相関ルール「X → I_j」の「サポート」と呼ぶ。また、トランザクション集合 T の全体に対する上記 K の割合をアイテムセット X の「サポート」と呼ぶ。

2

【０００３】アイテムセット X、アイテム I_j の組合せは多数ある得るが、その中から、与えられた最小コンフィデンス c、および最小サポート s 以上のコンフィデンスとサポートを持つ相関ルールを発見するための基本的な手法について文献「Fast Algorithms for Mining Association Rules」（Proceedings of VLDB、1994）に述べられている。この文献では、n 個のアイテムからなるアイテムセットのうち、最小サポート s 以上のサポートを持つものの集合をラージアイテム集合 L_n と呼び、L_(k-1) を元に L_k を得る処理を１つのパスとし、k の値を1 ずつ増加させながら、新たなラージアイテム集合が得られなくなるまでパスを繰り返すことによって最小サポート s を満たすアイテムセットの集合を求める。

【０００４】L_(k-1) を元に L_k を得るには、まず、L_(k-1) に含まれるうちの k 個のアイテムからなる可能な全てのアイテムセットを作成し、これらアイテムセットの集合を候補アイテム集合 C_k とし、次にトランザクション集合 T を走査し、C_k のそれぞれのアイテムセットについてアイテムセットを満足するトランザクションの数を数え上げ、それによってサポートの値を算出する。候補アイテム集合 C_k のうちで s 以上のサポートを持つアイテムセットの集合を新たなラージアイテム集合 L_k とする。この処理を、新たなラージアイテム集合が得られなくなるまでパスを繰り返す。ラージアイテム集合が得られた後、これに含まれるアイテムセットのそれぞれにおいて、含まれるアイテムを用いて作成可能な相関ルールについてそのコンフィデンスを算出し、最小コンフィデンス c を満たす相関ルールを選び出す。こうして得られた相関ルールが最終的な結果となる。

【０００５】上記の基本的アルゴリズムを計算機で実行する際には、トランザクションの集合を２次記憶に保持し、候補アイテム集合を満たすトランザクションの数を数え上げるカウンタを主記憶に保持することになる。この時、２つの問題点がある。１つは、１回のパスを実行するごとにトランザクション集合全体を走査するため、２次記憶からのデータの読み出しに多くの処理時間を費やしてしまうという点である。もう１つは、与えられたトランザクション集合に現れるアイテムの数によっては、候補アイテム集合が非常に大きくなり、カウンタを保持するために大容量の主記憶が必要となる点である。これらの問題を解決するための並列分散処理方法が文献「Parallel Mining of Association Rules」（IEEE Transactions on Knowledge and Data Engineering、1996）に述べられている。この文献には３つの並列アルゴ

リズムが説明されている。これら3つのアルゴリズムはいずれも、複数の処理装置を有し、トランザクション集合は分割されて各処理装置に局所的な2次記憶に分配されて保持されることを前提としている。第1の並列アルゴリズムは「Count Distribution」と呼ばれ、候補アイテム集合のサポートの算出の際のトランザクション集合の走査を複数の処理装置で並列に行うことによってデータの読み出しに要する時間を短縮することが可能である。（しかし、候補アイテム集合のサポートを算出する際のカウンタを各々の処理装置で全て保持するので、大容量の主記憶が要求されると言う点は解決されない。）第2の並列アルゴリズムは「Data Distribution」と呼ばれ、カウンタを複数の処理装置に分配して保持することによって、各処理装置において必要となる主記憶量を削減することが可能である。（しかし、各処理装置に分配されたトランザクションを処理装置間で転送する必要があるという別の問題が生じる。）また、Count Distribution、Data Distribution ともに、各々の処理装置で得られた候補アイテム集合のサポートを1回のパスごとに集計するので、処理装置ごとの処理時間に差がある場合、1回のパスごとに最も遅い処理装置の終了を待たねばならず、処理時間の無駄が生じるという問題もある。第3の並列アルゴリズムは「Candidate Distribution」と呼ばれ、この問題を解決することを目的としている。あるアイテムセットを満たすトランザクションの集合をサポートトランザクション集合と呼ぶことにする。

【0006】Candidate Distribution アルゴリズムでは途中のパスm（m はヒューリスティックに決定する）において、候補アイテム集合 C_m に含まれるアイテムセットをグループ分けする。この時、異なるグループに属するアイテムセットの間でサポートトランザクション集合がなるべく重ならないようにする。このグループ分けにしたがって候補アイテム集合と、そのサポートトランザクション集合を複数の処理装置へ分配し直す。これにより、各々の処理装置へ分配されたトランザクションのみの走査で以降のパスにおけるラージアイテム集合を得ることができ、全ての処理装置がパスごとに同期を取る必要がなくなる。

【0007】相関ルール以外のデータマイニング手法として、特徴ルール（CharacteristicRule）と、その発見法が文献「Characteristic Rule Induction Algorithm forData Mining」（proceedings of PAKDD、1998）に述べられている。この文献では、複数のフィールドからなるレコードの集合を分析対象データとしている。特徴ルールは「if A then B」と記述される。 A は1個以上の条件の組合せ、B は単一の条件である。ここで言う「条件」とは、フィールドとその値の組であり、例えば、「X_i」をフィールド、「v_ij」を値とすれば、これらを組にした物は条件であり、「X_i = v_ij」と記述される。また、分析対象のレコードにおいて、フィールド X_i に値 v_ij を持つ場合、そのレコードは条件「X_i = v_ij」を満足すると言う。また、特徴ルールは評価値を持ち、特徴ルール「if A thenB」の評価値は

【0008】

【数1】P(A)^a log{P(B¦A)／P(B)}

と定義される。ここで、P(A)、P(B) はそれぞれ、分析対象レコード全体のうちで条件 A および条件 B を満足するレコードの割合であり、 P(B¦A) は条件 Aを満足するレコードのうち、条件 A と 条件 B の両方を満足するレコードの割合である。また、指数 a はヒューリスティックに定められる正の定数である。

【0009】この文献で述べられている特徴ルールの発見法は、分析対象データと then 部の条件、if 部に含まれる条件数の上限、ルール数 M が与えられた時に、特徴ルールの生成と評価値の算出を繰り返し、評価値の最も大きい M 個の特徴ルールを発見する方法である。

【0010】上記の文献に述べられている特徴ルール発見法では、1つの特徴ルールを生成し、その評価値を算出するごとに分析対象データの全体、または一部を走査する必要があり、相関ルールの場合と同じように、2次記憶からのデータの読み出しに時間がかかるという問題がある。これを解決するために、分析対象データの走査を一回のみ行う方法が特開平11-3360号「大規模データ分析方法」である。この方法は、可能な全ての条件について、これを満足するレコードの数を数え上げるためのカウンタを用意し、一回のデータ走査で全ての条件についてのレコードの数え上げを行うというものである。これによれば、データの読み出しにかかる時間を削減することができる。しかし一方、カウンタを保持するために大容量の主記憶が必要になるという問題がある。

【0011】

【発明が解決しようとする課題】相関ルール、特徴ルールのいずれの場合も、基本的な発見方法においては、分析対象データの走査にかかる処理時間とカウンタ保持に必要な主記憶の容量の問題があり、一方を小さくしようとすると、他方が大きくなるという問題がある。これを解決するための並列分散処理方法がいくつかあるが、相関ルール発見の並列分散処理方法に関しては、データを分割して複数の処理装置に分配するので、分析対象データの走査にかかる処理時間は小さくできるものの、処理装置間で分析対象データの転送が必要となったり、カウンタを保持するために、依然として各々の処理装置において大容量の主記憶が必要となる。また、データマイニング機能を持たない通常のデータベースシステムに分析対象データが保持されている場合、並列分散処理方法を実行するためには、前処理としてデータベースから各処理装置へデータを分配する必要があり、このための処理時間が余計にかかるという問題がある。

【0012】本発明の目的は、複数の処理装置を用いた並列分散処理において、分析対象データの転送量、転送

回数を少なく抑え、かつ、各々の処理装置において必要となる主記憶の量を少なく抑え、かつ、通常のデータベースシステムに接続して実行可能なデータマイニングの方法を提供することである。

【００１３】

【課題を解決するための手段】本発明の並列分散分析方法では、単数、または複数のデータ格納手段と複数のデータ分析手段を用いる。単数のデータ格納手段を用いる場合、全ての分析対象データは単数のデータ格納手段に格納される。複数のデータ格納手段を用いる場合、分析対象データは分割され、複数のデータ格納手段に分配されて格納される。

【００１４】データ格納手段は分析対象データを複数のデータ分析手段に対して送信する。複数のデータ分析手段は同一のデータを受信し、受信したデータを対象としてそれぞれのデータ分析手段において分析を行う。その後、それぞれのデータ分析手段において得られた分析結果をまとめて、全体の分析結果とする。

【００１５】データ格納手段は分析対象データを複数のデータ分析手段に対してまとめて一回送信し、複数のデータ分析手段はこれを共有する。すなわち、データ格納手段は複数のデータ分析手段の各々に対して個別に分析対象データを送信することはしない。

【００１６】また、本発明の並列分散分析方法では、複数のデータ分析手段が共有する共有記憶手段を用いる場合がある。複数のデータ分析手段は、それぞれの分析結果を共有記憶手段上に保持する。したがって、共有記憶手段に保持された分析結果を読み出すことによって全体の分析結果を得ることができる。

【００１７】

【発明の実施の形態】本発明の第一の実施の形態を説明する。図１に本実施形態の構成を示す。本実施形態は、データ格納装置１０１、複数のデータ分析装置１０２、分析結果集計装置１０３がバス型通信路１０４によって接続されている。分析対象データはデータ格納装置に格納されている。

【００１８】図２にデータ分析の手順をしめす。データ分析装置の準備処理２０１では、データ分析装置のそれぞれにおいてデータ分析の準備を行い、分析対象データを受信する準備が完了したら、準備完了の信号をデータ格納装置へ送信する。データ格納装置では、全てのデータ分析装置からの準備完了信号を受信するのを待つ（処理２０２）。全てのデータ分析装置からの準備完了信号を受信後、処理２０３において、データ格納装置は分析対象データを走査し、まだ送信していないデータが残っている場合は処理２０４へ、残っていない場合（全ての分析対象データを送信し終わった場合）は処理２０６へ進む。レコード送信処理２０４では、データ格納装置は、まだ送信していないデータのうちの１レコードを送信する。

【００１９】一度送信されたレコードは送信済みとして扱われる。送信されたレコードはバス型通信路を介して複数のデータ分析装置へ伝送される。通信路がバス型であるため、レコードはデータ格納装置からの１回の送信で、通信路に接続された全てのデータ分析装置へ伝送される。データ受信、分析処理２０５では、データ格納装置から送信されたレコードをそれぞれのデータ分析装置において受信する。レコードの受信後、データ分析装置では次のレコードを受信する準備を行い、受信準備が完了したら、準備完了の信号をデータ格納装置へ送信する。また、既に受信済みのレコードとともに受信したレコードを対象としてデータ分析を行う。データ受信、分析処理２０５の後は処理２０２へ戻る。分析結果集計処理２０６では、分析結果集計装置はデータ分析装置から分析結果を受け取り、これを集計して全体の分析結果を得る。以上が、分析方法の概要である。このように、複数の装置が互いに協調しながら、データ分析を行う。以下に、特徴ルールの発見を例にとり、各装置において行われる処理を詳細に説明する。

【００２０】分析対象データは複数のフィールドからなるレコードの集合であり、全てのレコードは同数のフィールドを持つ。レコードは分析の対象となる対象物、フィールドは対象物の持つ属性に対応する。商店の顧客情報を例にとると、１つのレコードは１人の顧客、各フィールドは性別、年齢などの、顧客の属性に対応する。特徴ルール発見では前処理として、各属性値を少数のカテゴリに変換する。例えば、「年齢」は通常１０～１００程度の範囲の値を取り得るが、これを「３５歳以下」、「３６歳以上５５歳以下」、「５６歳以上」のような少数のカテゴリに変換する。また、「性別」は、もともと「男」「女」の２つの値しか取り得ないので、このまま２つのカテゴリとして用いることが多い。このようにカテゴリへの変換を施した分析対象データの例を図３に示す。

【００２１】特徴ルールは、次のように書き表される。

【００２２】「if 性別=男 and 年齢=56以上 then 購入額=大」すなわち、対象物の属性とカテゴリ化された値の組からなる if-then ルールである。特徴ルールの if 部に現れる属性を「条件項目」、 then 部に現れる属性を「結論項目」と呼ぶ。１つの属性が同時に条件項目と結論項目の両方になることはない。また、特徴ルールは評価値を持つ。一般に特徴ルールを「if A then B」と表した時、その評価値は次式で定義される。

【００２３】

【数２】$P(A)^a \log\{P(B|A)/P(B)\}$

ここで、P(A)、P(B) はそれぞれ、分析対象データ全体のうちで条件 A および条件 B を満足するレコードの割合であり、 P(B|A) は条件 A を満足するレコードのうち、条件 A と 条件 B の両方を満足するレコードの割合である。また、指数 a はヒューリスティックに定め

られる正の定数である。また、評価値の別の定義として、次の式を用いる場合もある。

【００２４】

【数３】P(A)^a P(B¦A)log{P(B¦A)／P(B)}

数１、数２のいずれにおいても、ルールに現れる条件を満たすレコード、および分析対象データ全体のレコードの数を知ることによって評価値を算出することができる。

【００２５】特徴ルール発見とは、上記で定義したルール評価値に基づき、評価値の大きい特徴ルールを発見する処理である。この時、発見すべき特徴ルールの数の上限、結論項目となるフィールドとその値、条件項目の候補となる複数のフィールド、１つの特徴ルールに現れる条件項目の数の上限が分析者によって与えられているものとする。図４、５、６にルール発見の手順を示す。図４はデータ格納装置における処理手順、図５はデータ分析装置における処理手順、図６は分析結果集計装置における処理手順である。分析対象データとして図３に示したデータを例にとり、結論項目を「購入額」、その値を「大」、条件項目の候補を「性別」、「年齢」、「職業」、条件項目の条件数の上限を２とする。また、「性別」は「男」、「女」の２値、「年齢」は「３５歳以下」、「３６歳から５５歳」、「５６歳以上」の３値、「職業」は「有」、「無」の２値を取り得るものとする。

【００２６】まず、データ格納装置、データ分析装置、分析結果集計装置の全てにおいて、割り当て設定処理４０１、５０１、６０１を行う。割り当て設定処理においては、指定された条件項目の候補と、条件数の上限にしたがって、可能な全ての特徴ルールに対応するカウンタを用意する。上記の条件項目候補と条件数の上限を用いた場合、条件項目の可能な組合せは２３通りであり、図７に示す２３通りの特徴ルールが可能である。特徴ルールの評価値算出には、前述の P(A)、P(B¦A)、P(B) が必要である。ここで、結論項目となるフィールドとその値は１つに指定されているため、 P(B) は全ての特徴ルールについて同じである。したがって、各特徴ルールについて、 P(A)、P(B¦A) を知るための２つのカウンタを用意することになる。これらのカウンタは複数のデータ分析装置に分配して割り当てられる。

【００２７】どのデータ分析装置にどの特徴ルールのカウンタを割り当てるかはデータ分析装置のうちの１つ、または、分析結果集計装置、またはデータ格納装置のいずれかによって決定される。そして、カウンタを割り当てられたデータ分析装置のそれぞれの識別名、または、カウンタを割り当てられたデータ分析装置の数がデータ格納装置へ通知される。各データ分析装置では割り当てにしたがってカウンタを用意する。

【００２８】データ格納装置では割り当て設定処理４０１の後、カウンタを割り当てられた全てのデータ分析装置と分析結果集計装置から、準備完了の信号を受信するのを待ち（処理４０２）、準備完了の信号を全て受信したら、処理４０３へ進む。処理４０３では、未送信のデータがあるかどうかを確認し、未送信のデータがある場合はレコード送信処理４０４へ、全てのデータが送信済みである場合は送信終了処理４０５へ進む。レコード送信処理４０４では、まだ送信していないデータのうちの１個のレコードを通信路へ送信し、処理４０２へ戻る。送信終了処理４０５では、全てのデータを送信し終わったことを示す信号を通信路へ送信する。

【００２９】以上で、データ格納装置における処理は終了する。

【００３０】データ分析装置では割り当て設定処理５０１の後、分析対象データを受信する準備が完了したことを示す信号をデータ格納装置へ送信する（処理５０２）。データ受信待ち処理５０３では、データ格納装置からデータを受信するのを待ち、データを受信したら、処理５０４へ進む。処理５０４においては、受信したデータがデータの終了を示す信号であるかどうかを判定し、データ終了信号であればルール評価処理５０７へ、そうでなければカウンタ更新処理５０５へ進む。カウンタ更新処理５０５では、受信したデータは分析対象データのレコードであるとして、そのフィールドの値を評価する。そして、フィールドの値が、データ分析装置に割り当てられた特徴ルールの条件と一致していれば、該当するカウンタの値を更新する。一般に、１個のレコードは、複数の特徴ルールの条件と一致し得る。すなわち、１個のレコードの処理において、複数の特徴ルールのカウンタ更新が行われる。データ受信準備処理５０６では、処理５０３で受信したレコードを破棄し、次のレコードを受信する準備をし、処理５０２へ戻る。ルール評価処理５０７では、データ分析装置に割り当てられたルールの評価値をカウンタの値に基づいて算出し、評価値の大きい順に、発見すべき特徴ルールの数の上限として指定された数の特徴ルールを取り出す。ただし、評価値が０よりも小さい特徴ルールは取り出されない。したがって、指定された数よりも少数の特徴ルールしか取り出されない場合がある。この後、分析結果集計装置からの指示を待つ処理５０８へ進む。分析結果集計装置からの指示が終了指示である場合（処理５０９）は、データ分析装置における処理を終了する。分析結果集計装置からの指示が分析結果送信指示である場合（処理５１０）は、取り出しておいた特徴ルールとその評価値を分析結果集計装置へ送信し（処理５１１）、指示を待つ処理５０８へ戻る。

【００３１】分析結果集計装置では割り当て設定処理６０１の後、分析対象データを受信する準備が完了したことを示す信号をデータ格納装置へ送信する（処理６０２）。データ受信待ち処理６０３では、データ格納装置からデータを受信するのを待ち、データを受信したら、

処理６０４へ進む。処理６０４においては、受信したデータがデータの終了を示す信号であるかどうかを判定し、データ終了信号であれば分析結果収集処理６０６へ、そうでなければ受信準備処理６０５へ進む。受信準備処理６０５では、データ格納装置から受信したデータを破棄し、次のデータを受信する準備をし、処理６０２へ戻る。分析結果収集処理６０６では、全てのデータ分析装置へ順に分析結果送信指示を送り、それぞれのデータ分析装置で取り出された特徴ルールとその評価値を収集する。分析結果集計処理６０７では、収集した特徴ルールの中から、評価値の大きい順に、発見すべき特徴ルールの数の上限として指定された数の特徴ルールを取り出し、これを特徴ルール発見の結果とする。終了指示処理６０８では、全てのデータ分析装置へ終了指示を送信する。以上で、分析結果集計装置における処理を終了する。

【００３２】図８に本発明の第二の実施の形態の構成を示す。第一の形態ではバス型の通信路を使用していたが、第二の形態ではリング型の通信路を使用している。リング型通信路では、全ての装置は送信端子と受信端子を持ち、装置の送信端子は別の装置の受信端子と単方向の通信路を介して接続されている。データ格納装置、データ分析装置、分析結果集計装置はいずれも、データを発信する場合は、発信する装置の識別子をデータに付加して、送信端子からデータを送出する。データを受信する場合は、受信端子からデータを受け取る。そして、そのデータに付加された識別子が自分のものであれば、そのデータを破棄し、識別子が自分のものでなければ、必要に応じてそのデータを装置内に取り込むとともに、データと識別子を自分の送信端子から送出する。このように、ある装置から送信されたデータは全ての装置の間を順に転送されて送信元の装置へ戻り、そこで転送が終了する。

【００３３】したがって、バス型の通信路と同様に、送信元の装置からの１回の送信で、他の全ての装置にデータが伝送される。

【００３４】図９に本発明の第三の実施の形態の構成を示す。第二の形態ではスター型の通信路を使用している。スター型通信路では、全ての装置は双方向の通信路を介して集線装置９０１と接続されている。集線装置は複数の接続端子を持ち、１つの接続端子において受信した信号を他の全ての接続端子から送信する機能を持つ。したがって、バス型の通信路と同様に、送信元の装置からの１回の送信で、他の全ての装置にデータが伝送される。

【００３５】図１０に本発明の第四の実施の形態の構成を示す。データ管理装置１００４、複数のデータ分析装置１００２、分析結果集計装置１００３が１つの共有メモリ１００５に接続され、データ管理装置１００４は通信路を介してデータ格納装置１００１に接続されている。分析対象データはデータ格納装置に格納されている。

【００３６】共有メモリは、データ区画１００６、カウンタ区画１００７、分析結果区画１００８に分けられている。

【００３７】図１１、１２、１３に本実施形態における特徴ルール発見の手順を示す。図１１はデータ管理装置における手順、図１２はデータ分析装置における手順、図１３は分析結果集計装置における手順である。まず、データ管理装置、データ分析装置、分析結果集計装置の全てにおいて、割り当て設定処理１１０１、１２０１、１３０１を行う。割り当て設定処理においては、指定された条件項目の候補と、条件数の上限にしたがって、可能な全ての特徴ルールに対応するカウンタを共有メモリのカウンタ区画内に用意する。また、特徴ルールのそれぞれについて、そのカウンタ更新処理を担当するデータ分析装置を決める。どのデータ分析装置にどの特徴ルールのカウンタ更新処理を割り当てるかはデータ分析装置のうちの１つ、または、分析結果集計装置のいずれかによって決定される。また、データ分析装置のそれぞれの分析結果を書き込む領域を分析結果区画に用意する。また、データ管理装置において、共有メモリのデータ区画内に複数のレコードバッファ１００９を用意する。１つのレコードバッファは１つのレコードを格納できるレコード領域と、カウンタ更新処理を割り当てられた複数のデータ分析装置の１つ１つと対応するフラグ領域からなる。フラグの１つ１つは「データ無効」、「データ有効」のどちらかの状態を取る。１つのレコードバッファのフラグが全て「データ無効」である場合、そのレコードバッファは「空いている」と言う。初期状態では、全てのレコードバッファは空いている。また、レコードバッファとは別に、データ区画内にデータ終了フラグ１０１０を用意する。データ終了フラグは「真」、「偽」の２つの状態の一方を取ることができ、初期状態は「偽」である。

【００３８】データ管理装置では、データ格納装置から分析対象データのレコードを１個入力する（処理１１０２）。この時、データ格納装置からデータの終了を示す信号を受信した場合は処理１１０６へ、そうでなければ処理１１０４へ進む（処理１１０３）。処理１１０４では、共有メモリのレコードバッファを走査し、空いているレコードバッファを１つ探し出す。これが見つからなかった場合は、見つかるまで走査を繰り返す。空いているレコードバッファが見つかった場合は処理１１０５へ進む。処理１１０５では、処理１１０４で見つかったレコードバッファのレコード領域にデータ格納装置から入力したレコードを格納し、そのレコードバッファの全てのフラグに「データ有効」を示す値を設定する。その後、処理１１０２へ戻る。処理１１０６では、データ区画のデータ終了フラグを「真」の状態にする。以上でデ

ータ管理装置における処理を終了する。

【0039】データ分析装置では、共有メモリのレコードバッファを走査し、自データ分析装置に対応するフラグが「データ有効」であるレコードバッファを1つ探す（処理1202）。これが見つかった場合は処理1203へ、見つからなかった場合は処理1205へ進む。処理1203では、処理1202で見つかったレコードバッファからレコードを読み込み、そのフィールドの値が、データ分析装置に割り当てられた特徴ルールの条件と一致していれば、該当する特徴ルールのカウンタの値を更新する。処理1204では、処理1202で見つかったレコードバッファ内の、自データ分析装置に対応するフラグを「データ無効」に更新する。その後、処理1202へ戻る。処理1205では、共有メモリ内のデータ終了フラグを調べ、状態が「真」であれば処理1206へ進み、状態が「偽」であれば処理1202へ戻る。処理1206では、データ分析装置に割り当てられたルールの評価値をカウンタの値に基づいて算出し、評価値の大きい順に、発見すべき特徴ルールの数の上限として指定された数の特徴ルールを取り出す。ただし、評価値が0よりも小さい特徴ルールは取り出されない。処理1207では、取り出した特徴ルールとその評価値を共有メモリの分析結果領域に書き込む。以上でデータ分析装置における処理を終了する。

【0040】分析結果集計装置では、共有メモリの分析結果領域を調べ、全てのデータ分析装置の分析結果が分析結果領域に書き込まれるのを待つ（処理1302）。全てのデータ分析装置の分析結果が揃ったら、分析結果領域に書き込まれた特徴ルールの中から、評価値の大きい順に、発見すべき特徴ルールの数の上限として指定された数の特徴ルールを取り出し、これを特徴ルール発見の結果とする（処理1303）。以上で分析結果集計装置における処理を終了する。

【0041】なお、以上で説明した第四の実施の形態はカウンタを共有メモリ内に置く形態であったが、データ分析装置のそれぞれが局所メモリ1401を持っている場合は、カウンタを局所メモリ内に置くこともできる（図14）。

【0042】

【発明の効果】本発明によれば、大量のデータを多面的に分析するデータマイニングなどのデータ分析処理を複数の処理装置を並列に動作させて実行する場合に、分析対象データの転送量、転送回数を少なく抑え、かつ、各々の処理装置において必要となる主記憶の量を少なく抑えることができる。また、大量データ分析方法に適するように設計されたデータ格納方法を特に必要としないので、一般のデータベースシステムに格納されたデータを分析対象とすることができる。

【図面の簡単な説明】

【図1】第1の実施形態の構成図。

【図2】分析方法の概要を示す流れ図。

【図3】分析対象データの例を示す図。

【図4】第1の実施形態のデータ格納装置における処理を示す流れ図。

【図5】第1の実施形態のデータ分析装置における処理を示す流れ図。

【図6】第1の実施形態の分析結果集計装置における処理を示す流れ図。

【図7】全ての可能な特徴ルールを列挙した図。

【図8】第2の実施形態の構成図。

【図9】第3の実施形態の構成図。

【図10】第4の実施形態の構成図。

【図11】第4の実施形態のデータ管理装置における処理を示す流れ図。

【図12】第4の実施形態のデータ分析装置における処理を示す流れ図。

【図13】第4の実施形態の分析結果集計装置における処理を示す流れ図。

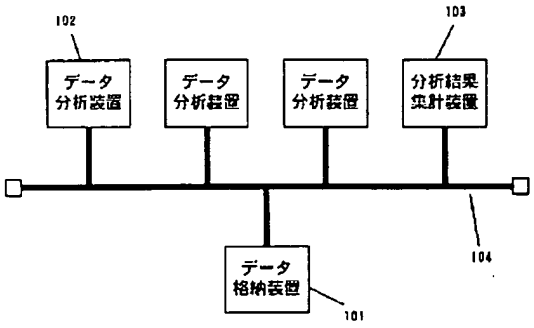
【図14】第4の実施形態において、局所メモリにカウンタを置く構成図。

【符号の説明】

101…データ格納装置、102…データ分析装置、103…分析結果集計装置、104…バス型通信路、1004…データ管理装置、1005…共有メモリ、1006…共有メモリ内のデータ区画、1007…共有メモリ内のカウンタ区画、1008…共有メモリ内の分析結果区画、1009…レコードバッファ、1010…終了フラグ、1401…局所メモリ内のカウンタ。

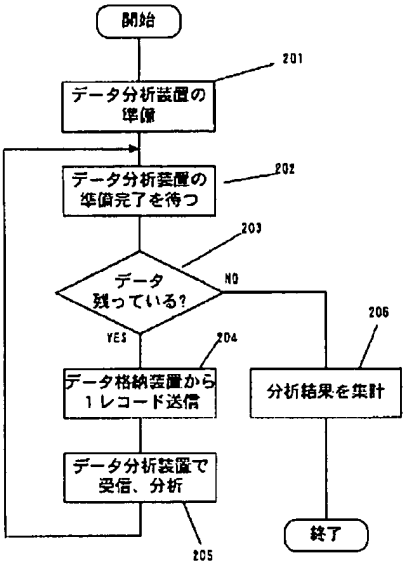
【図1】

図 1



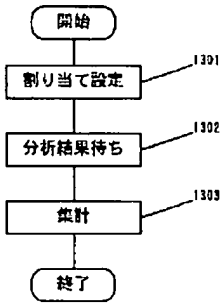
【図2】

図 2



【図13】

図 13



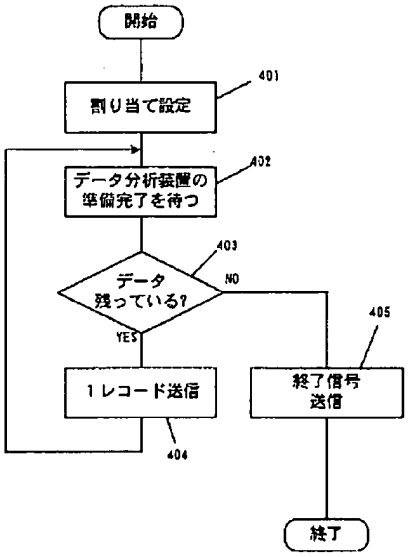
【図3】

図 3

顧客ID	性別	年齢	職業	購入額
0011	男	～35	無	小
0013	男	36～55	有	大
0132	女	36～55	有	大
0322	男	56～	有	中
0022	女	～35	有	大
0051	女	～35	無	中
..	..	..	..	..

【図4】

図 4



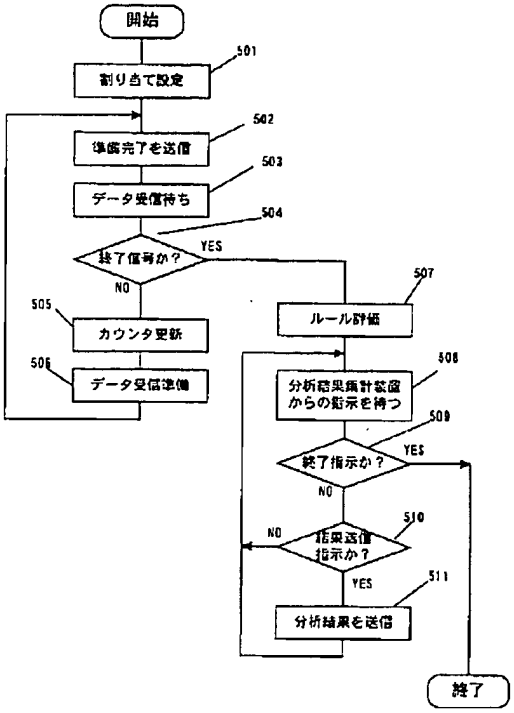
【図7】

図 7

```
if 性別＝男                          then 購入額＝大
if 性別＝男 and 年齢＝35以下         then 購入額＝大
if 性別＝男 and 年齢＝36～55         then 購入額＝大
if 性別＝男 and 年齢＝56以上         then 購入額＝大
if 性別＝男 and 職業＝有             then 購入額＝大
if 性別＝男 and 職業＝無             then 購入額＝大

if 性別＝女                          then 購入額＝大
if 性別＝女 and 年齢＝35以下         then 購入額＝大
if 性別＝女 and 年齢＝36～55         then 購入額＝大
if 性別＝女 and 年齢＝56以上         then 購入額＝大
if 性別＝女 and 職業＝有             then 購入額＝大
if 性別＝女 and 職業＝無             then 購入額＝大

if 年齢＝35以下                      then 購入額＝大
if 年齢＝35以下 and 職業＝有         then 購入額＝大
if 年齢＝35以下 and 職業＝無         then 購入額＝大

if 年齢＝36～55                      then 購入額＝大
if 年齢＝36～55 and 職業＝有         then 購入額＝大
if 年齢＝36～55 and 職業＝無         then 購入額＝大

if 年齢＝56以上                      then 購入額＝大
if 年齢＝56以上 and 職業＝有         then 購入額＝大
if 年齢＝56以上 and 職業＝無         then 購入額＝大

if 職業＝有                          then 購入額＝大
if 職業＝無                          then 購入額＝大
```

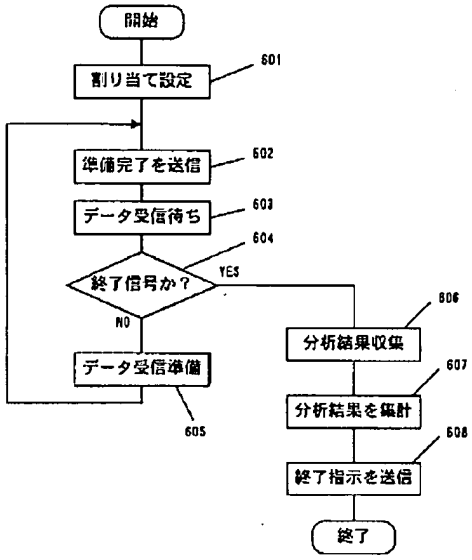
【図5】

図 5



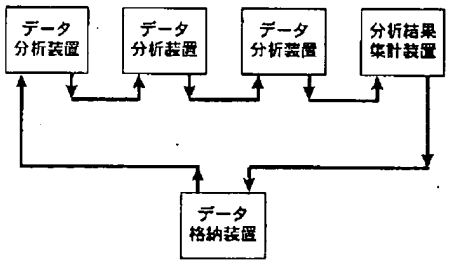
【図6】

図 6



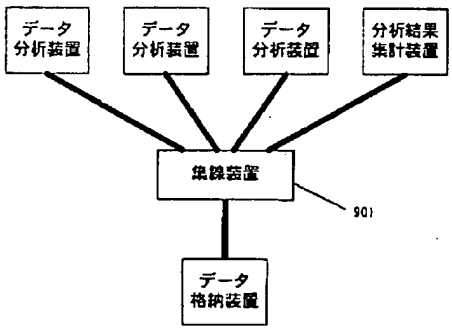
【図8】

図 8



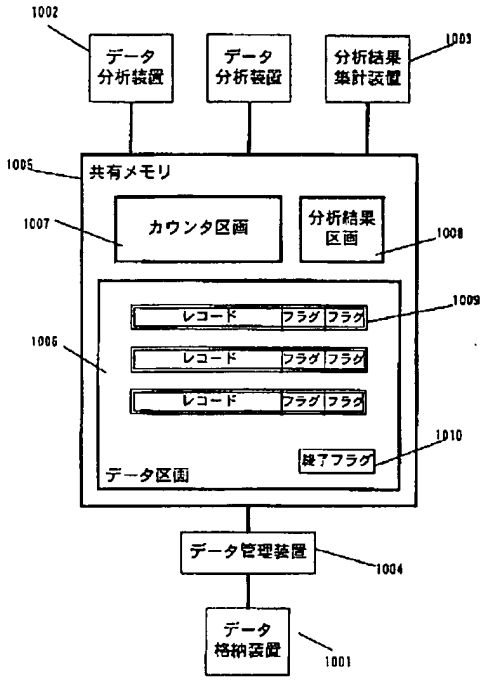
【図9】

図 9



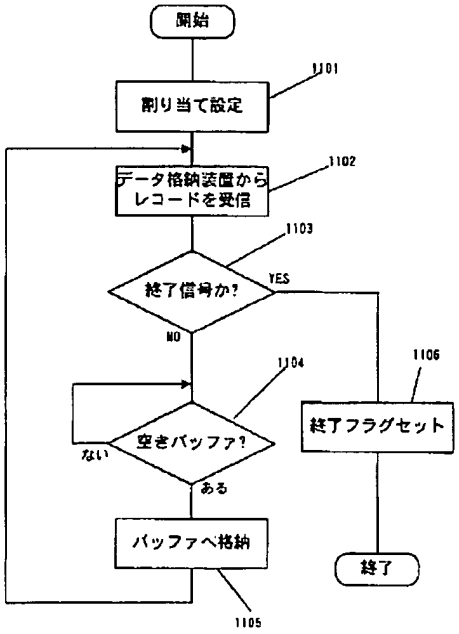
【図10】

図 10



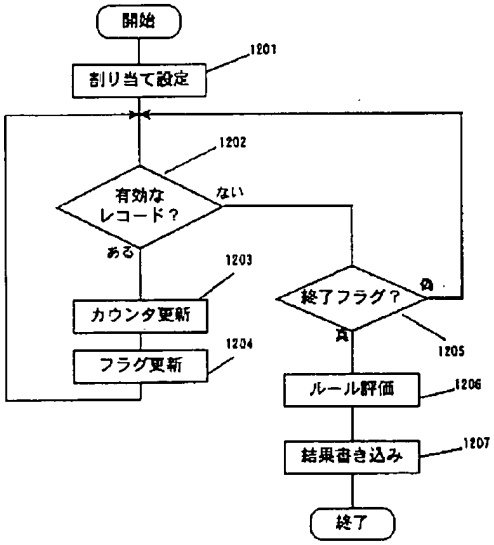
【図11】

図 11



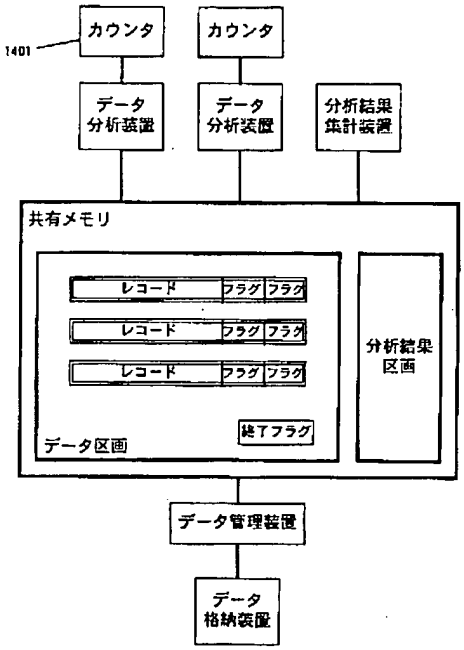
【図12】

図 12



【図14】

図 14



フロントページの続き

(72)発明者　伊藤　幸康
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア事業部内

Fターム(参考)　5B075 KK02 PQ05 QS05

PAT-NO: JP02001167098A

DOCUMENT-IDENTIFIER: JP 2001167098 A

TITLE: DISTRIBUTED PARALLEL ANALYZING METHOD FOR MASS DATA

PUBN-DATE: June 22, 2001

INVENTOR-INFORMATION:

NAME	COUNTRY
MAKI, HIDEYUKI	N/A
MORITA, TOYOHISA	N/A
ITO, YUKIYASU	N/A

ASSIGNEE-INFORMATION:

NAME	COUNTRY
HITACHI LTD	N/A

APPL-NO: JP11347026

APPL-DATE: December 7, 1999

INT-CL (IPC): G06F017/30

ABSTRACT:

PROBLEM TO BE SOLVED: To solve the problems of necessity to transfer analysis object data between processors for dividing data and distributing them to plural processors, to distribute data from a data base to each of processors as preprocessing for executing a parallel distributed processing method and to require a bulk main memory for the processor conventionally concerning the parallel distributed processing method of data mining for finding knowledge out of mass data.

SOLUTION: This method uses a single or plural data storage means, an analyzed result accumulating means and plural data analyzing means. The data storage means transmits analysis object data once and

these data are received by the plural data analyzing means. Each of data analyzing means performs analysis with the received data as an object and afterwards, the analyzed results provided in the respective data analyzing means are accumulated by the analyzed result accumulating means and defined as the entire analyzed result.